XIOS benchmarking and profiling roadmap + preliminary results

Xavier Yepes-Arbós

Mario C. Acosta

The research leading to these results has received funding from the EU H2020 Framework Programme under grant agreement no. 823988
This material reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains

10/12/20

XIOS current developments

HPC I/O challenges for Earth System Modelling

- Exascale supercomputers will allow to make simulations at an unprecedented level of horizontal resolution.

- ...but this has implications:
  - A huge amount of data will be generated that must be efficiently written into the storage system.
  - No more offline post-processing is affordable due to the size of the "raw" data.

HPC I/O challenges for Earth System Modelling

- From a computational point of view, XIOS is thought to address:
  - The inefficient legacy read/write process.
  - The unmanageable size of “raw” data.

- By implementing:
  - Scalable parallel I/O.
  - Online post-processing.

- ...but it has been only tested for petascale supercomputers, so it is necessary to:
  - Design a battery of benchmarks to be as close as possible to future exascale machines.
  - Stress different aspects such as memory consumption, MPI scalability, netCDF parallel I/O or netCDF compression.

What will it be necessary from the computational side of XIOS?

- Extreme scalability on:
  - Memory use.
  - Model MPI processes (XIOS clients).
  - XIOS servers.
  - Writing into the storage system (very dependent on the file system).
- Efficient management of:
  - Memory.
  - Data affinity: optimal transfer from clients to servers (and two-level servers).
- Efficient online post-processing:
  - A good trade-off between size reduction and computational cost for the compression filter.
  - Exploit all the available computational resources to speed up costly filters: OpenMP and OpenACC.
  - Consider porting all filters from client to server side to not limit the model scalability

Questions to be investigated

- How will XIOS scale in exascale supercomputers?

- How will XIOS fit into nodes if memory management does not scale?

- How will XIOS deal with huge volumes of data without efficient compression?

- Will “one_file” mode (parallel writing) work for a lot of writers?

Planned XIOS benchmarking at BSC

- These are the models, versions and configurations that will be used:
  - OpenIFS 43r3: Tco1279L137 (9 km).
  - EC-Earth 4 (both OpenIFS and NEMO components using XIOS at the same time): 10 km demonstrator from WP1.
  - XIOS trunk.
- Strategy:
  - Perform all the benchmarks with OpenIFS standalone and, if needed, extrapolate them to EC-Earth 4.
  - Use Yann’s toy model to reproduce issues found in OpenIFS to work on them easily. The toy model can be run on other machines.
- HPC cluster: MareNostrum 4
  - Computing nodes: 48 cores, 96 GB of main memory and 100 Gbit/s Intel Omni-Path.
  - GPFS file system.

Contributions from other institutions

- Cerfacs offered to test the interpolation performance and quality for different pair of grids (as they already did for the SCRIP and for ESMF) in Yann’s toy environment.

- CMCC offered to perform the benchmarking in the CMCC Zeus machine of the critical parts by using the toy model framework available in XIOS.

Output scalability benchmarks

Scale XIOS servers and memory consumption according to these conditions:

- Use two different I/O loads:
  - Real output configuration similar to CMIP6 experiments.
  - Very large output configuration (theoretical experiment with many 3D fields).
- Fix a large amount of MPI processes for the model, at least 5000:
  - Scale number of nodes for XIOS servers.
  - Scale number of XIOS servers.
- Fix an amount of XIOS servers and XIOS exclusive nodes to scale XIOS clients.
- Use both one_file and multiple_file modes.
- Use both one or two level servers (ratio, pools, timeseries).
- Test the affinity of XIOS servers to reduce data movement.
- Local storage strategy: Write chunked netCDF files directly to the local scratch of the XIOS servers nodes.

Post-processing cost benchmarks

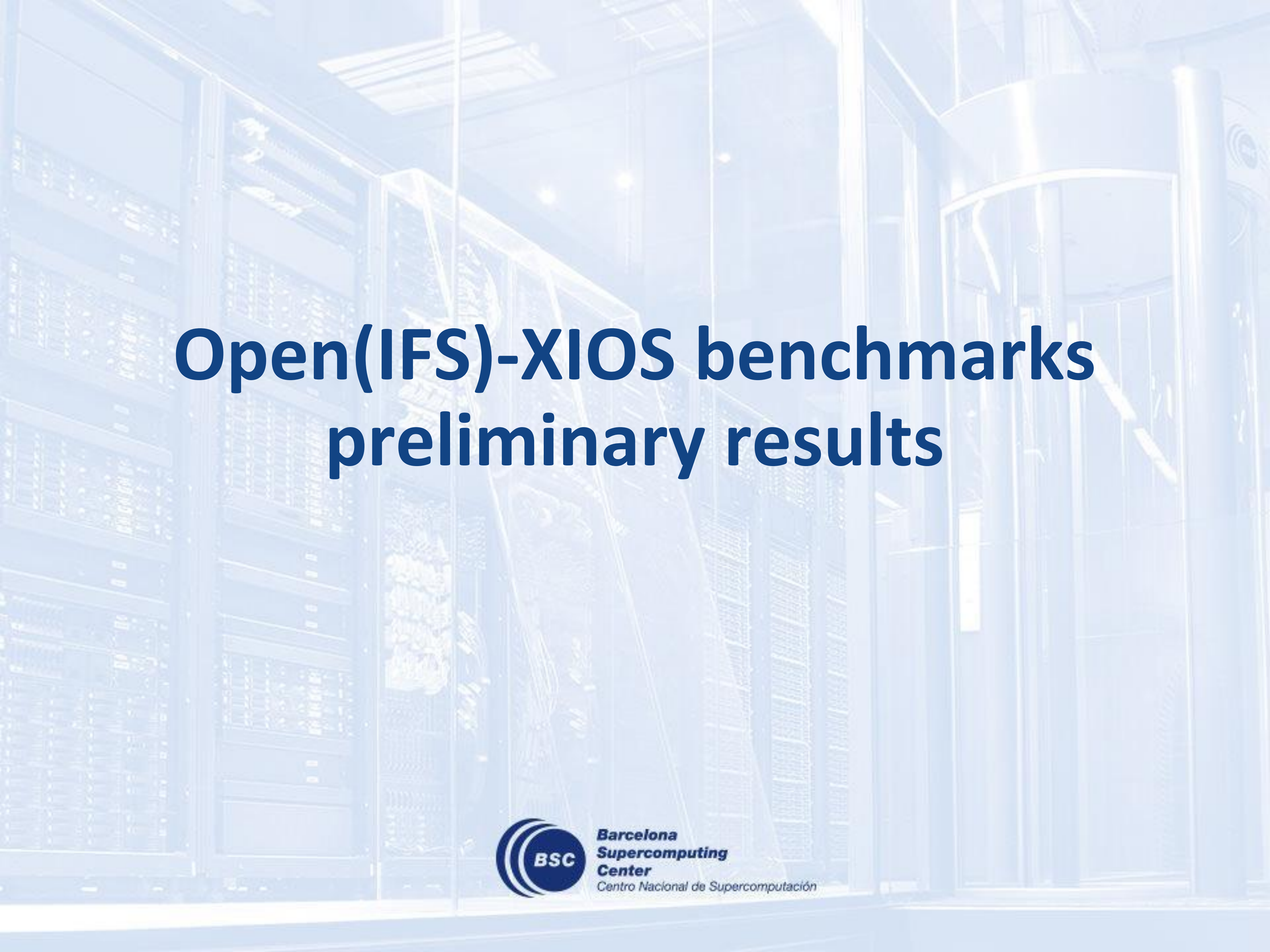
Determine the cost associated to perform common filters:

- Spatial filters (e.g. horizontal interpolation from unstructured to regular).
- Temporal filters (e.g. daily average).
- Arithmetic filters (e.g. compute wind speed and direction from u and v components).
- Compression filter.
  - In the future explore the new parallel compression of HDF5.

Profiling

Make use of advanced performance analysis tools such as Extrae and Paraver:

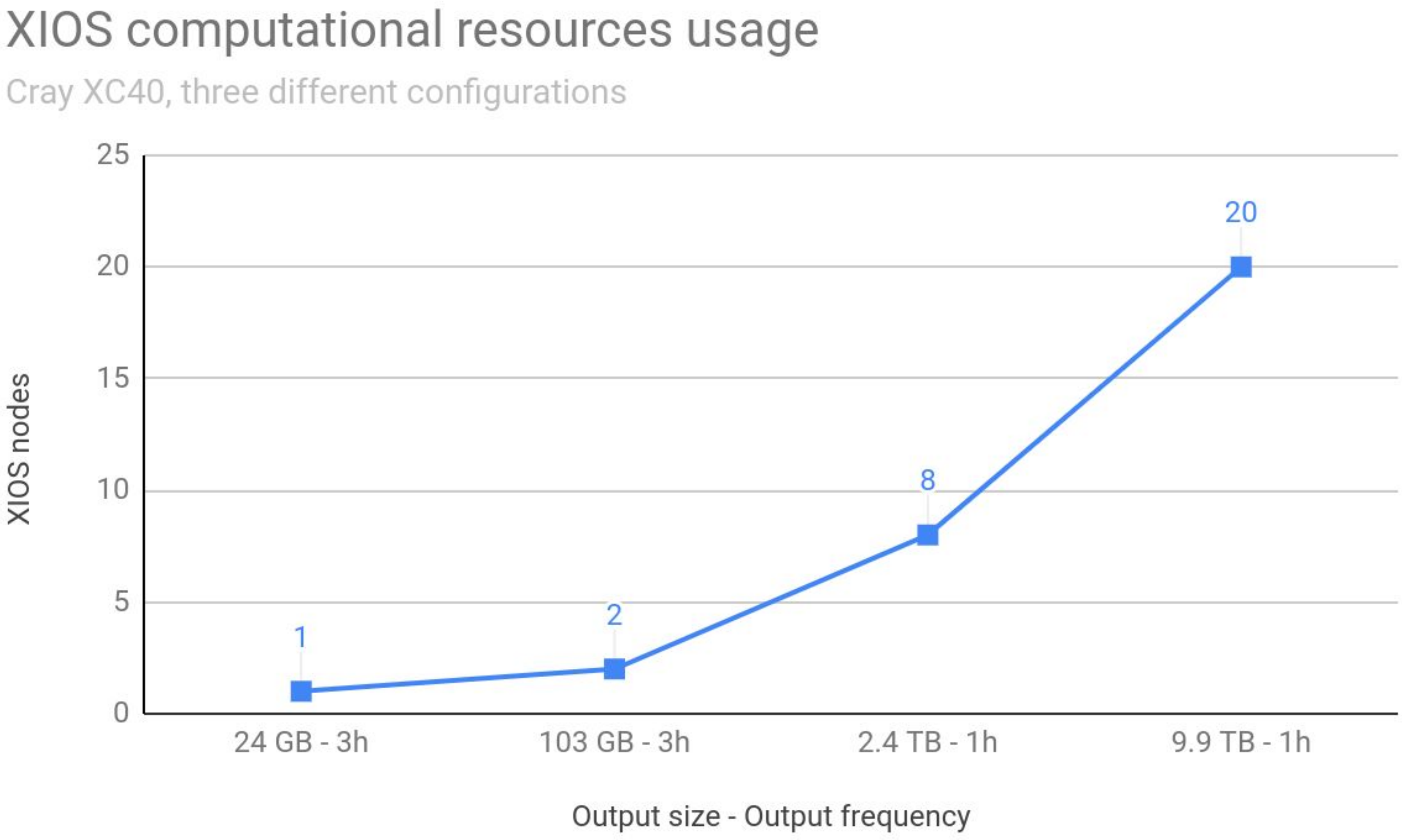
- Basic performance analysis.
- Analyze both the computational performance and memory consumption of netCDF parallel writing (one_file mode) and netCDF sequential writing (multiple_file mode).
- Analyze common filters.
- If it is possible, compare the XIOS performance in a GPFS or a Lustre filesystem.



Open(IFS)-XIOS benchmarks preliminary results

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

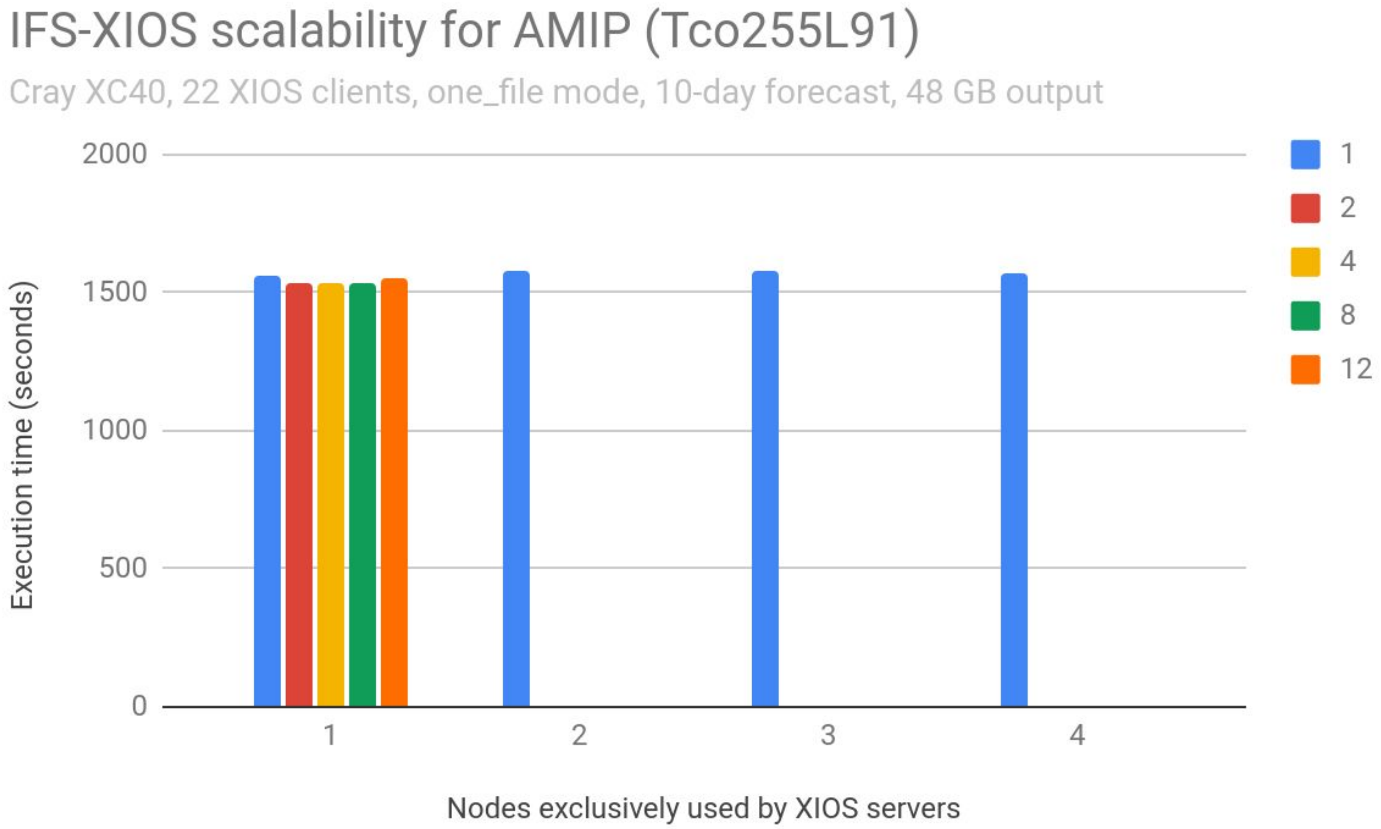
XIOS servers resources usage



Preliminary XIOS servers scalability

Note 1: the legend indicates the number of XIOS servers per node.

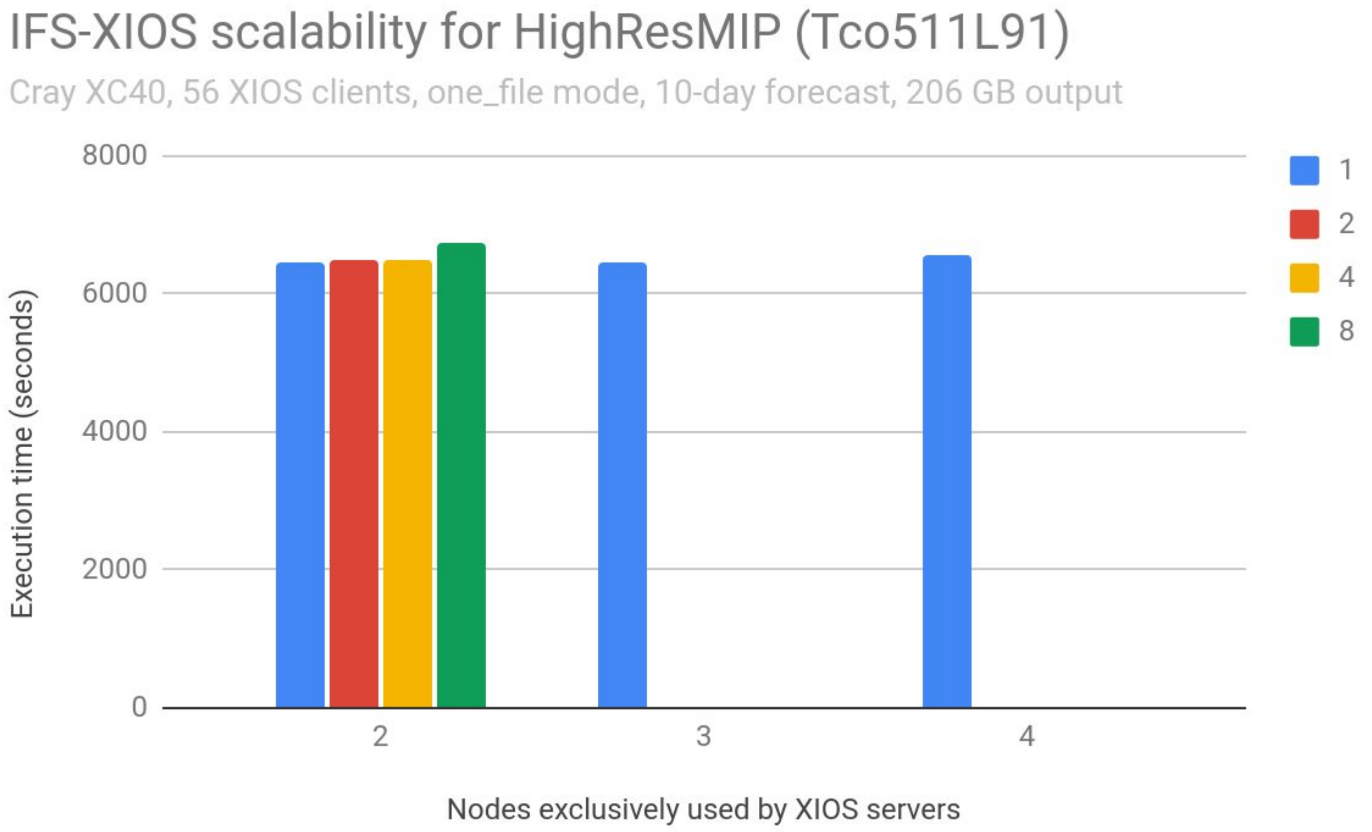
Note 2: the lowest number of nodes (x axis) also indicates the minimum number of nodes that XIOS requires to run due to its memory consumption.



Preliminary XIOS servers scalability

Note 1: the legend indicates the number of XIOS servers per node.

Note 2: the lowest number of nodes (x axis) also indicates the minimum number of nodes that XIOS requires to run due to its memory consumption.

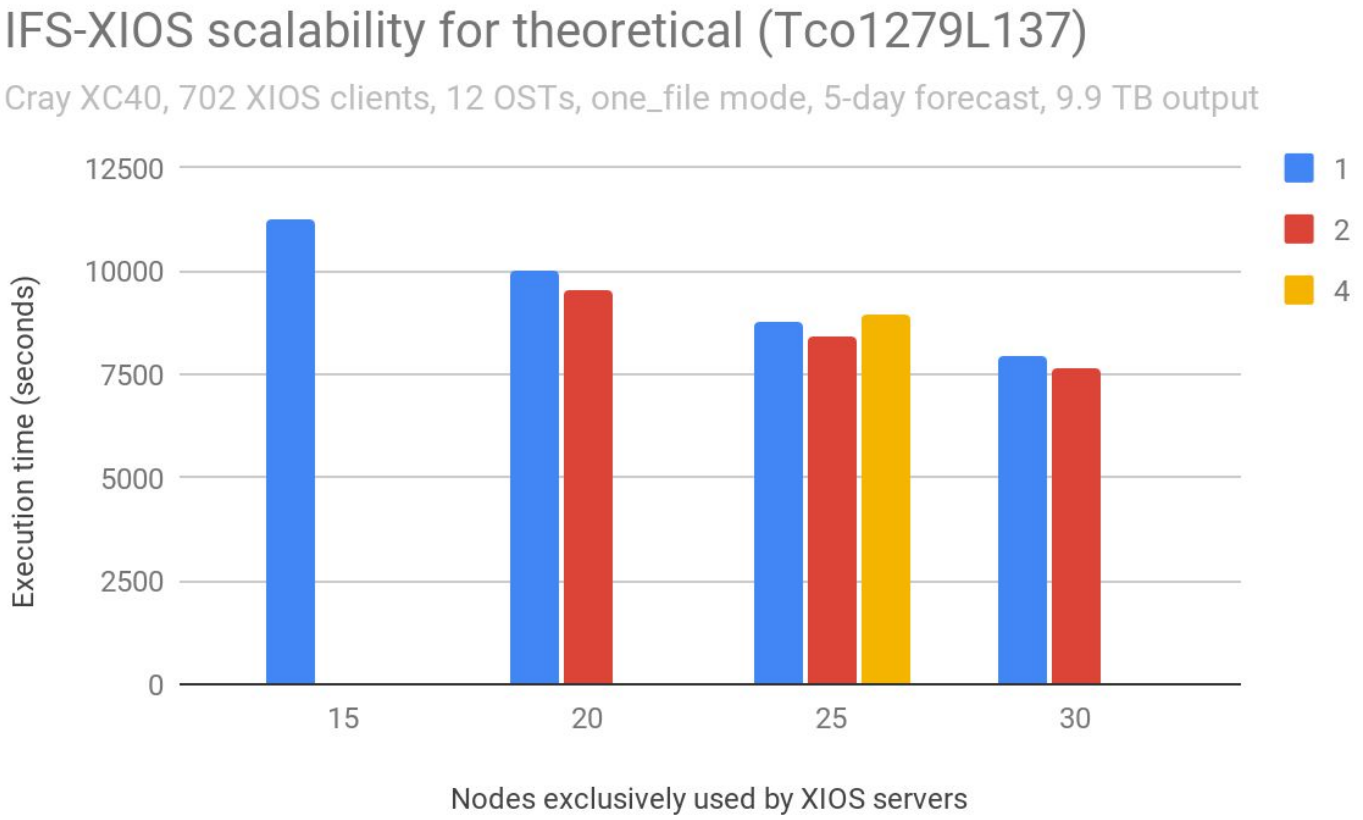


IFS-XIOS scalability for HighResMIP (Tco511L91)

Cray XC40, 56 XIOS clients, one_file mode, 10-day forecast, 206 GB output

Preliminary XIOS servers scalability

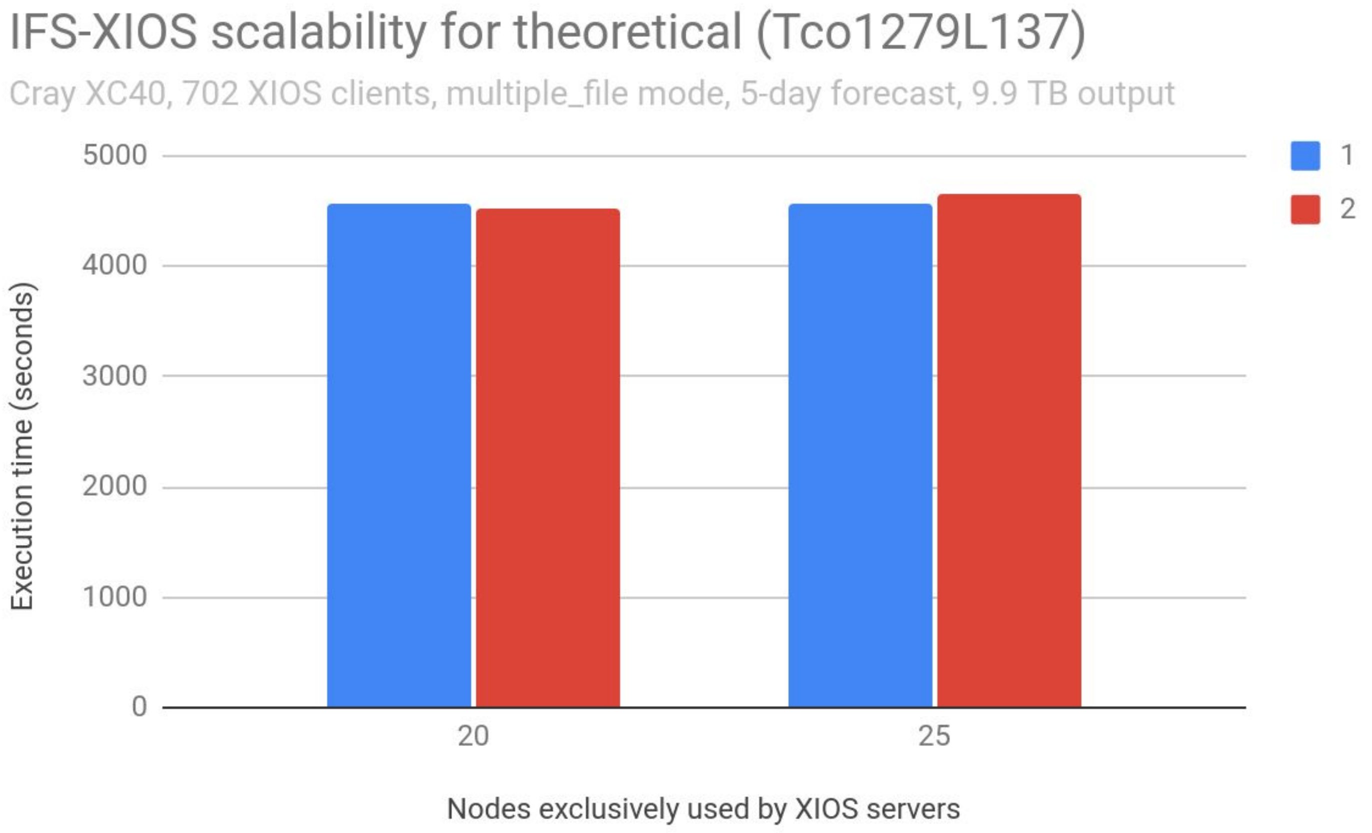
Note 1: the legend indicates the number of XIOS servers per node.
Note 2: the lowest number of nodes (x axis) also indicates the minimum number of nodes that XIOS requires to run due to its memory consumption.



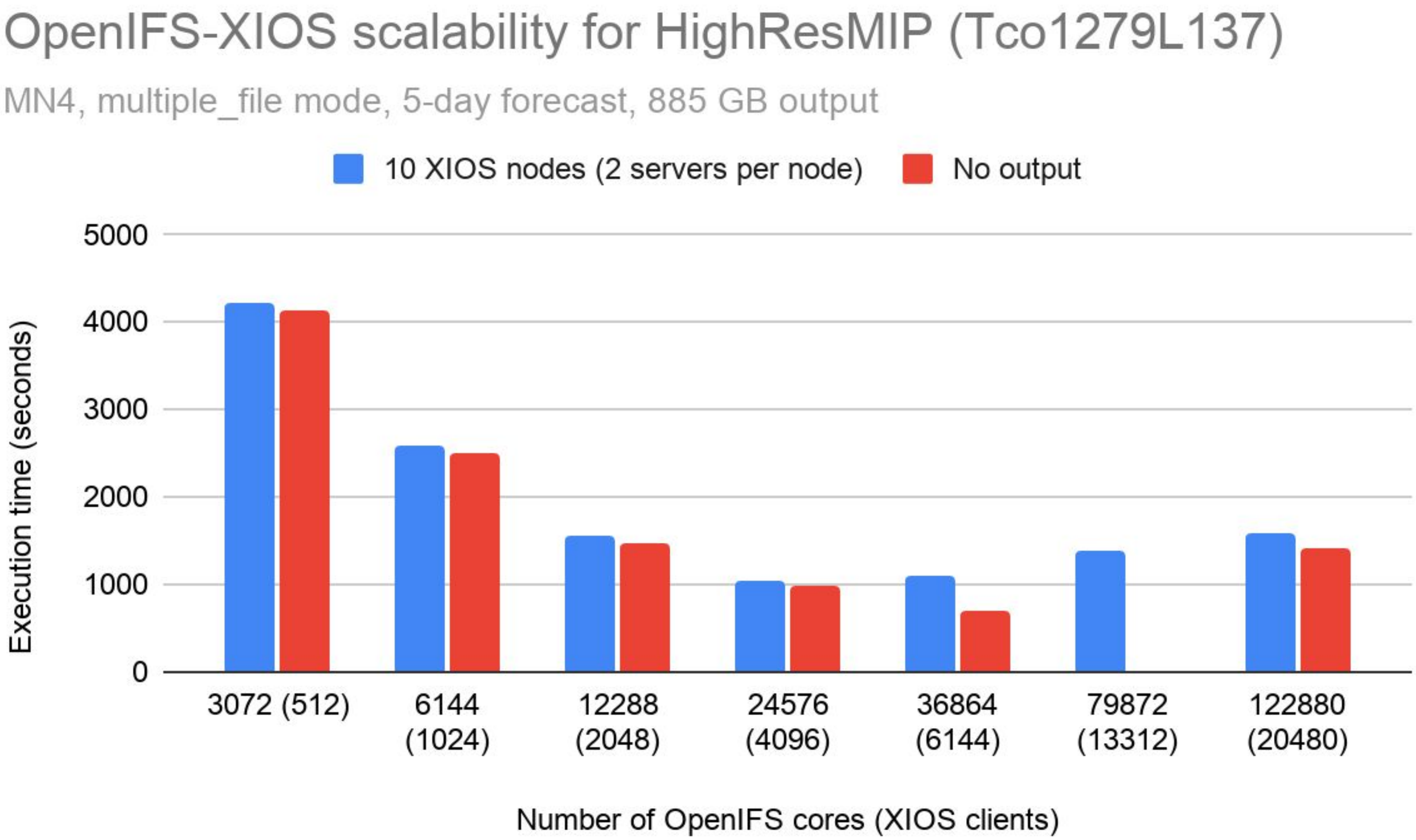
Preliminary XIOS servers scalability

Note 1: the legend indicates the number of XIOS servers per node.

Note 2: the lowest number of nodes (x axis) also indicates the minimum number of nodes that XIOS requires to run due to its memory consumption.



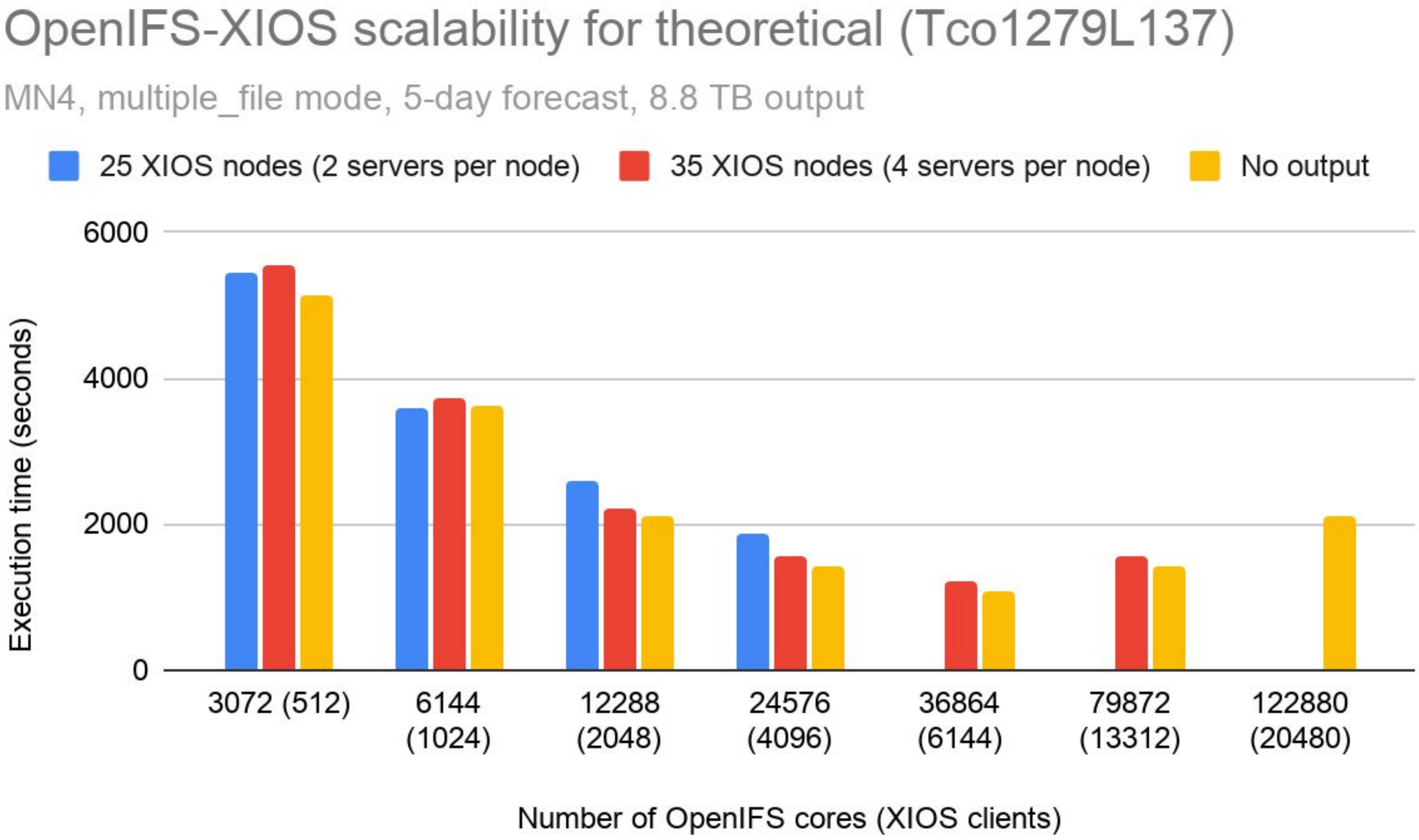
Preliminary XIOS clients scalability



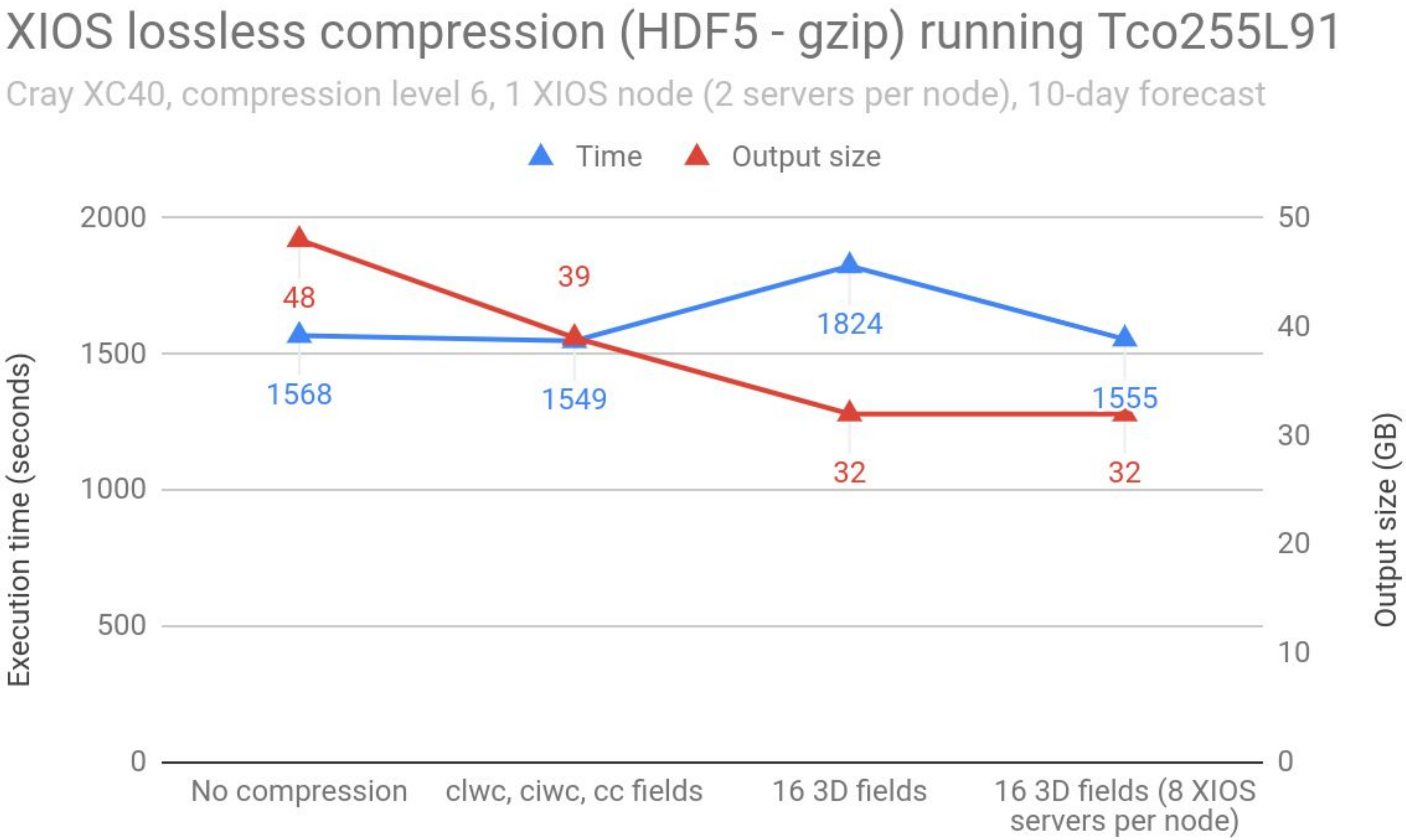
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Preliminary XIOS clients scalability



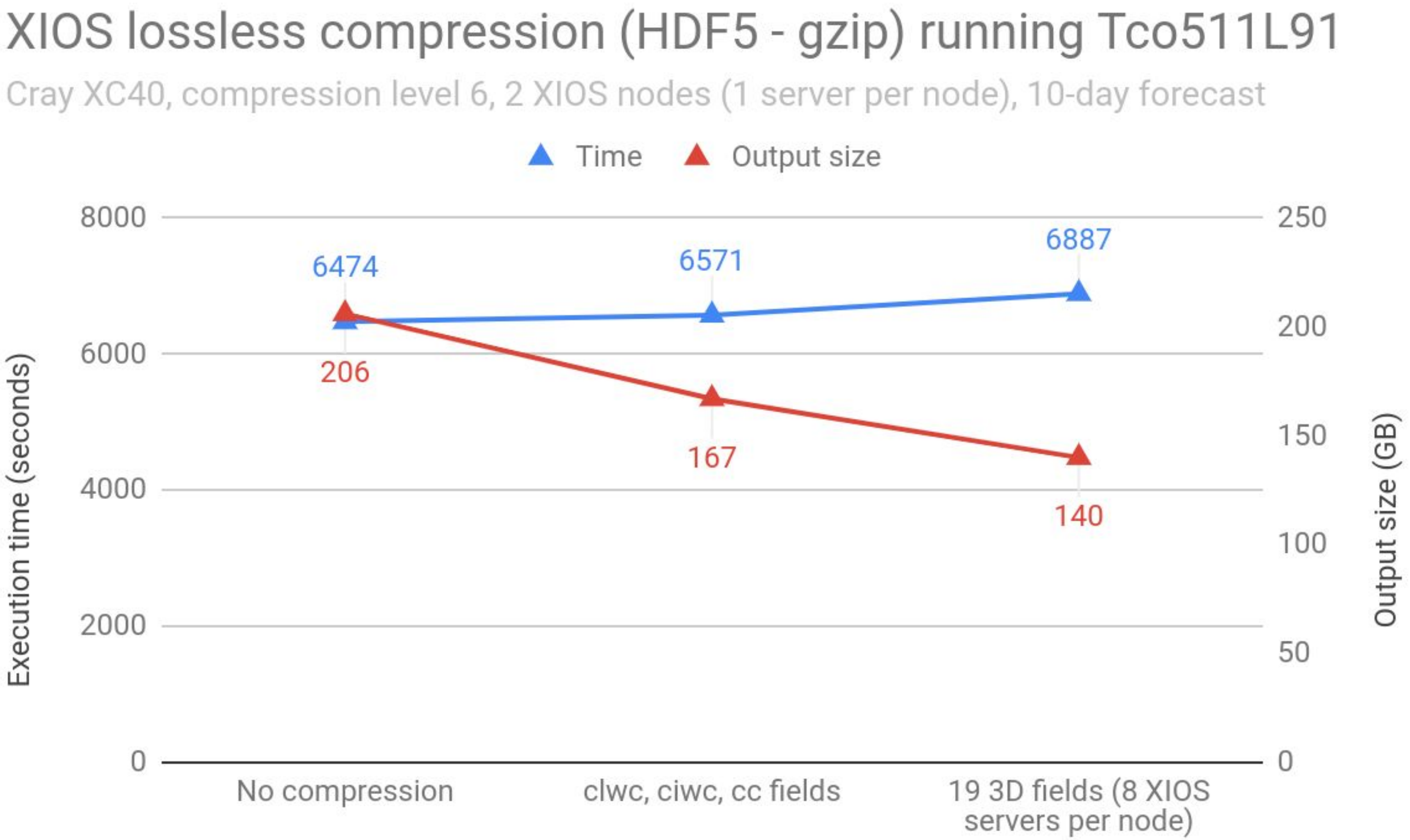
Default XIOS compression



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

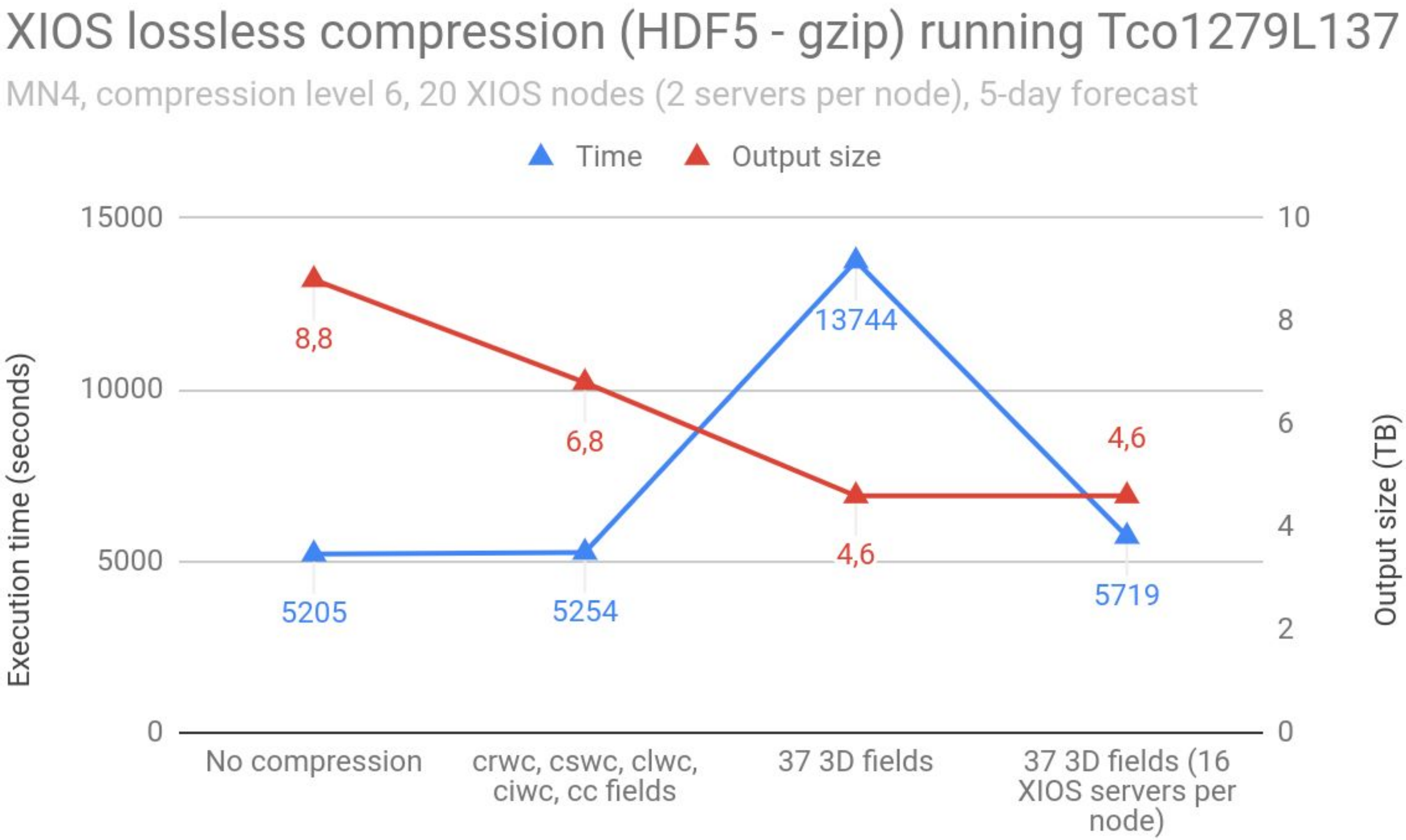
Default XIOS compression



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Default XIOS compression



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Preliminary conclusions

- One_file mode does not scale well.
  - HDF5 parallel I/O is very system dependent.
- XIOS servers consume a lot of memory although post-processing is disabled.
  - They depend on the output volume as well as output frequency.
- XIOS clients seem to scale well, but when there are many of them, they might add some overhead that can be mitigated adding more XIOS nodes for servers (to reduce the XIOS_clients/XIOS_servers_nodes ratio).
- XIOS compression through HDF5 (gzip) does not provide a good trade-off between high compression ratio and compression speed.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Thank you

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

xavier.yepes@bsc.es